Predicting U.S. Hospital Ratings with Performance Data

By: Franco Menares Paredes, Fangzhou Yuan, Xiyi(Britney) Wang, Jiangyue Zhu
Course: 36-600, Overview of Statistical Learning and Modeling

Introduction

This study examines a range of factors affecting hospital ratings, particularly emphasizing the importance of medical costs and the implementation of safety measures. By analyzing a comprehensive dataset of American hospital evaluations, we aim to create an analytical model capable of predicting hospital ratings.

The goal of this project is to find the optimal model that accurately classifies high-rated and low-rated hospitals, directing improvements in healthcare quality and patient satisfaction.

Data

The dataset is comprised of detailed records from 1,739 hospitals across the United States. It features 20 predictive variables that inform the binary `Rating` response variable, which classifies hospital performance as "low" (1-3 stars) or "high" (4-5 stars). The predictors consist of both numerical and categorical variables, encompassing hospital performance metrics and costs associated with key procedures.

Variable	Definition	Variable	Definition
Facility.Type	Type of hospital	Procedure.Heart.Attack.Value	Value for heart attack care
Rating.Mortality	Mortality rate comparison	Procedure.Heart.Failure.Cost	Cost for heart failure care
Rating.Safety	Safety comparison	Procedure.Heart.Failure.Quality	Heart failure care quality
Rating.Readmission	Readmission rate comparison	Procedure.Heart.Failure.Value	Value for heart failure care
Rating.Experience	Patient experience comparison	Procedure.Pneumonia.Cost	Cost for pneumonia care
Rating.Effectiveness	Effectiveness of care comparison	Procedure.Pneumonia.Quality	Pneumonia care quality
Rating.Timeliness	Timeliness of care comparison	Procedure.Pneumonia.Value	Value for pneumonia care
Rating.Imaging	Use of imaging comparison	Procedure.Hip.Knee.Cost	Cost for hip/knee care
Procedure.Heart.Attack.Cost	Cost for heart attack care	Procedure.Hip.Knee.Quality	Hip/knee care quality
Procedure.Heart.Attack.Quality	Heart attack care quality	Procedure.Hip.Knee.Value	Value for hip/knee care

Table 1: Predictor Variables

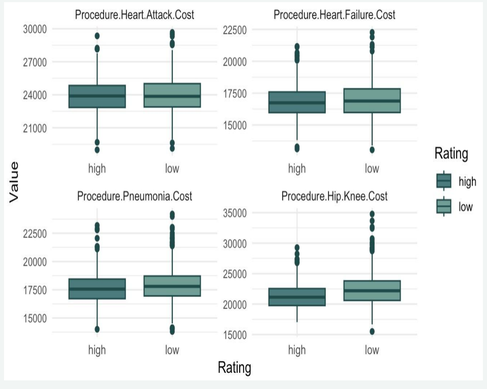

Figure 1: Box plots of quantitative variables

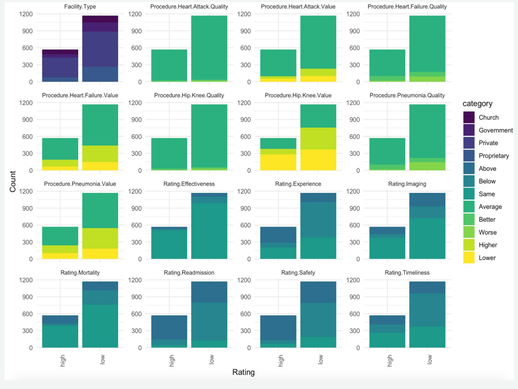

Figure 2: Bar plots of categorical variables

Modeling Analysis

- We splitted our data so that 80% is used for training and 20% for testing.
- We develop binary classification using logistic regression, linear-kernel SVM, decision tree, random forest, and XGBoost.
- For each of these models, we generate ROC curves for each model and compute the AUC values. The model with the highest AUC will be considered the best.

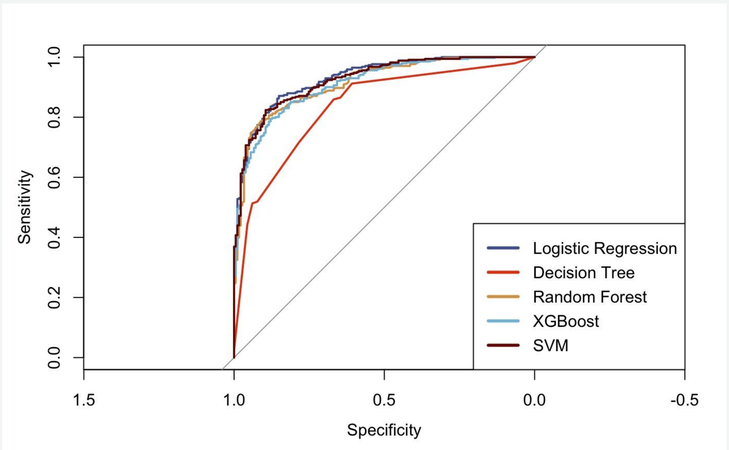

Figure 3: ROC curves of different models

Model	Logistic Regression	SVM	Decision Tree	Random Forest	XGBoost
AUC	0.930	0.925	0.835	0.914	0.912

Table 2: AUCs for each model

	Response = high	Response = low
Prediction = high	154	44
Prediction = low	27	297

Table 3: Confusion Matrix for Logistic Regression

Observations:

- The highest AUC 0.930 is from logistic regression model. The optimum threshold is 0.581, MCR is 0.136.
- Optimal class separations for all models are generated by maximizing Youden's J statistic, hence balancing the classification results and make better predictions for the class with a smaller sample size
- `Rating.Experience, Rating.Safety, Rating.Readmission` have the highest importance, followed by `Rating.Mortality, Procedure.Hip.Knee.Cost,Procedure.Pneumonia.Quality,` etc. Therefore, patient experience, hospital safety and readmission are most crucial for hospital ratings.

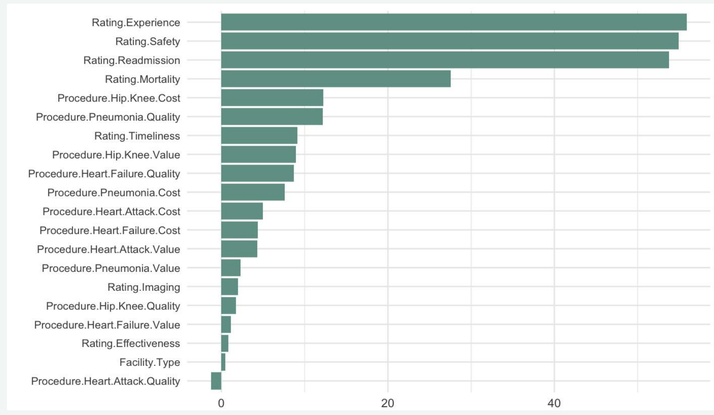

Figure 4: Feature Importance by Random Forest

Conclusion

Our analysis successfully applied various predictive models to determine the best approach for classifying U.S. hospitals into "high" and "low" rating categories based on a range of performance metrics. The logistic regression model demonstrated the highest efficiency with an AUC of 0.930, indicating a strong predictive capability.

This study lays the groundwork for future research where these models can be further refined and potentially applied in real-world scenarios to guide improvements in healthcare quality and patient satisfaction. Future analysis may focus on integrating additional variables, exploring alternative modeling techniques, or applying the model to different subsets of the healthcare system.